IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

Phishing Website and Spam Content Detection using Machine Learning Algorithms

Tanishka Narang[1], Snehal Kamble[2], Pallavi Sadavarte[3], Shivani Paygude[4], Prof. Anjali P. Kadam[5]
Students, Dept. of Computer Engineering[1,2,3,4]
Guide, Dept. of Computer Engineering[5]
Bharati Vidyapeeth's College of Engineering for Women, Pune, India

Abstract: *Phishing attacks continue to pose a major threat for computer system defenders, often forming the first step in a multi-stage attack. There have been great strides made in phishing detection; however, some phishing emails appear to pass through filters by making simple structural and semantic changes to the messages. In this paper, a Phishing and Spam Content Detection System is proposed that deals with the data uncertainty to which we are applying the SVM and NLP algorithm. There have been great strides made in phishing detection; however, some phishing URLs appear to pass through filters by making simple structural and semantic changes to the spellings. The phishing problem is big and there does not exist only one solution to reduce all susceptibilities effectively, thus multiple techniques are implemented. We can reduce the threat of phishing by analyzing various features of URL, then by checking the legitimacy of the website by knowing where the website is being hosted and who is managing it, another approach is to check visual appearance to analyze the genuineness of the website. The next step is to make sure the content on the analyzed website is spam or not. By using Natural Language Processing we process the content present on the website and determine whether it is spam or not. We make use of Machine Learning techniques and algorithms for the evaluation of these different features of URLs and websites. Using different approaches can improve the accuracy and enhance the system, thus helping better detect and prevent these threats.*

Keywords: Detection, Phishing, Spam, Support Vector Machine, Natural Language Processing

I. INTRODUCTION

Phishing can be defined as impersonating a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc. Phishing fraud might be the most widespread cybercrime used today. There are countless domains where phishing attacks can occur like the online payment sector, webmail, financial institution, file hosting or cloud storage, and many others. The webmail and online payment sector were embattled by phishing more than in any other industry sector.

Phishing can be done through email phishing scams and spear phishing hence users should be aware of the consequences and should not give their 100 percent trust in the common security applications. Machine Learning is one of the efficient techniques to detect phishing as it removes drawbacks of the existing approach.

II. LITERATURE SURVEY

1. Vaibhav Patil; Pritesh Thakkar; Chirag Shah; Tushar Bhat; S. P. Godse detected the problem of Web Spoofing which lures the user to interact with the fake websites rather than the real ones. To minimize this problem only one solution could not be used effectively, therefore multiple techniques were applied. In this paper, they have used 3 approaches to solve this problem. They analyzed various features of URLs then checked the legitimacy of the website by knowing where and by whom these websites were been hosted and managed and lastly checked the visual appearance and genuineness of the website. Implementing a hybrid solution would solve this problem which used 3 approaches.

2. Awishkar Ghimire, Avinash Kumar Jha, Surendrabikram Thapa, Sushruti Mishra, and Aryan Mani Jha presented different ways in which phishing URLs can be detected using a machine learning algorithm. Trained Naïve Bayes classifier, SVM, Regression trees and KNN. By experiment, regression trees gave the highest accuracy. Collected 450,176 URLs out of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

which 345,738 (77%) were benign websites and 104,438 (23%) malicious websites. 21 features were extracted and used for research. 5 Machine Learning algorithms and 5 ensemble algorithms were used on unbalanced, under-sampled, and over-sampled data. The classifiers were used and the validation was done using 10 fold cross-validation technique.

3. Pradeepthi K V, Kannan A applied pattern recognition capabilities of Machine Learning to the phishing detection domain. 4500 URLs out of which 2500 were genuine URLs and 2000 were phishing URLs and several classification algorithms were taken. Provided survey and comparative analysis of the different algorithms and identified which algorithm is best suited for detection. Detection was done just by analyzing the URL structure like no data entering and clicking.

4. Weiheng Bai analyzed structural features of the URL of the phishing website, extracted 12 features, and used 4 Machine learning algorithms for training. Used the best-performing algorithm as a model to identify unknown URLs. 7058 websites were taken out of which 3547 were malicious web pages taken from PhishTank and 3511 were benign pages taken from the Dmoz directory.

Features collected were: URL length, number symbol, a domain name with IP address, number of @ in URL, whether there is a symbol <//>, number of </> in URL, number of subdomains, length of a domain name, path length, HTTPS protocol, URL word segmentation features, and hist information characteristics. For feature evaluation, 6 features with the best classification effect were taken.

5. Mahajan Mayuri Vilas; Kakade Prachi Ghansham; Sawant Purva Jaypralash; Pawar Shila proposed that the motive of this study was to perform ELM derived from different 30 main components which were categorized using the Machine Learning approach. They used three ways for the detection of website phishing. The primitive approach evaluated different items of URL, the second approach analyzed the authority of a website and calculated whether the website is introduced or not and it also analyzed who is supervising it, the third approach checked the genuineness of the website.

6. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, and Dr. Aram Alsedrani collected 16000 URLs (legitimate and phishing) of 10 daily users out of which 12000 were phishing URLs and 4000 were legitimate URLs collected from PhishTank. Considered 36 features where 3 features were new. Categorized them into 3 main categories which were: features extracted from URL, features extracted from page content, and features extracted from page rank. The main function of this system was to decide the state of the website whether it is a phishing or legitimate website.

III. PROBLEM STATEMENT

Phishing Websites are duplicate web pages created to mimic real websites to deceive people. This project aims to provide a better detection rate in phishing website algorithms with spam detection.

IV. PROPOSED SYSTEM

In this project, we are using SVM and NLP algorithms when we are using the SVM algorithm then its accuracy is best. In this Project, we have provided URL and Phishing website data as Input. Then we detect the output whether the provided input is a Phishing website or not and whether the text data is spam or not.

V. SYSTEM ARCHITECTURE

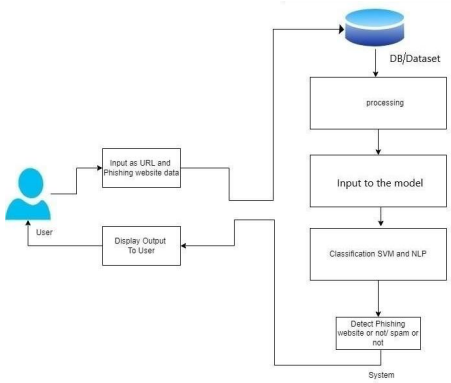

Fig 1. System Architecture

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

VI. METHODOLOGY

Here we have proposed a system for phishing website and spam content detection which includes various features and machine learning algorithms:

- New User Registration - User Registration is required to create new Login.
- New Login - The User has to Login after registering.
- Phishing Detection - A website URL is taken to detect whether it is a phishing website or not.
- Spam Detection - Website Content is accepted to detect as spam content or normal sentence.

VII. ALGORITHM

Support Vector Machine, or SVM, is a well-known Supervised Learning algorithm that is used for both classification and regression. However, it is mostly used in Machine Learning for Classification problems. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may simply place fresh data points in the proper category in the future.

A hyperplane is the optimal choice boundary. Linear SVM is used for linearly separable data, which implies that if a dataset can be categorized into two classes using a single straight line, it is considered linearly separable data, and the classifier employed is the Linear SVM classifier.

Non-linear SVM: Non-linear SVM is used for non-linearly separated data, which implies that if a dataset cannot be categorized using a straight line, it is used for non-linearly separated data.

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that helps machines to understand human language. Its goal is to build models that can make sense of the text provided and automatically execute tasks like translation, spelling check, or topic classification.

NLP examines the grammatical structure of sentences and the individual meaning of words then uses algorithms to extract meaning from them. That is to say, it makes sense of human language so that it can automatically perform different tasks. SVM and NLP are better than most of the other algorithms used as they have better accuracy.

SVM is a very good algorithm for classification. It's a supervised learning algorithm that is mainly used to classify data into different classes. SVM trains on a set of labeled data. The main benefit of SVM is that it can be used for both classification and regression.

The Random Forest (RF) and Support Vector Machines (SVM) were the machine learning model used, with the highest accuracies of 90% and 95% respectively. From the results obtained, the SVM is a better model than RandomForest in terms of accuracy. There are many algorithms used for classification in machine learning but SVM is better than most of the other algorithms used as it has better accuracy in results. This means that training an SVM will be longer to train than an RF when the size of the training data is higher. This has to be considered when choosing the algorithm. Space of the decision boundary separating the two classes that can also perform in n-Dimensional space.

VIII. RESULTS

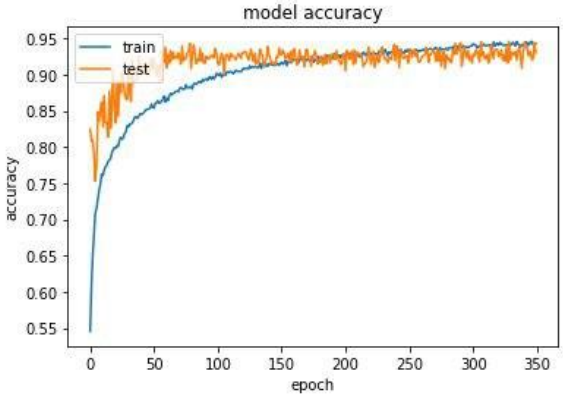

Fig 2. Model Accuracy

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

The proposed system accepts website URLs as input and displays whether they are phishing websites or not. It also accepts website content and provides the detection of the same as normal content or spam content. It has to be accessed through a complete registration and login system which provide a secure environment for the system. Using Support Vector Machine and Natural Language Processing techniques helps in providing better accuracy. The overall accuracy of this proposed system is about 85-88%.

IX. CONCLUSION

Phishing has become a serious network security problem, causing financial loss to both consumers and e-commerce companies. In this paper we've discussed implementing a system, using Machine Learning algorithms (SVM and NLP) for the prevention and detection of the same.

X. FUTURE SCOPE

Among several machine learning algorithms, SVM and NLP give better results. This work becomes unique from other existing work by proposing a group of features that can be extracted automatically using our own software tool. In the future, we can make the system available on mobile devices.

REFERENCES

[1]. Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat, and S. P. Godse, "Detection and Prevention of Phishing Websites using Machine Learning Approach", ICCUBEA 2018.

[2]. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, and Aram Alsedrani, "Detecting Phishing Websites Using Machine Learning", ICCAIS 2019.

[3]. Awishkar Ghimire, Avinash Kumar Jha, Surendrabikram Thapa, Sushruti Mishra, and Aryan Mani Jha, "Machine Learning Approach Based on Hybrid Features for Detection of Phishing URLs", ICoCCDSE 2021

[4]. Pradeepthi. K V and Kannan. A "Performance Study of Classification Techniques for Phishing URL Detection", ICoAC 2014

[5]. Weiheng Bai, "Phishing Website Detection Based on Machine Learning Algorithm",CDS 2020

[6]. Mahajan Mayuri Vilas, Kakade Prachi Ghansham, Sawant Purva Jaypralash, and Pawar Shila, "Detection of Phishing Website Using Machine Learning Approach", ICEECCOT 2019